General Bayesian Inference over the Stiefel Manifold via the Givens Transform

Arya A Pourzanjani Richard M Jiang Brian Mitchell Paul J Atzberger Linda R Petzold
University of California Santa Barbara

Abstract

We introduce an approach based on Givens Transforms to map between the space of orthonormal matrices and unconstrained parameters. Our approach allows for the application of general Bayesian inference algorithms to probabilistic models containing constrained unit-vectors or orthonormal matrix parameters. This includes a variety of matrix factorizations and dimensionality reduction models such as Probabilistic PCA (PPCA), Exponential Family PPCA (BXPCA), and Canonical Correlation Analysis (CCA). While previous Bayesian approaches to these models relied on separate sampling update rules for constrained and unconstrained parameters, our Givens Transform approach enables the treatment in many cases of unit-vectors and orthonormal matrices agnostically as unconstrained parameters. Thus Bayesian inference algorithms can be used on many of these models without modification. This opens the door to not just sampling algorithms but also Variational Inference (VI) methods. We illustrate with several examples and supplied code how our Givens Transform approach allows end-users to easily build complex models in their favorite Bayesian modeling framework such as Stan, Edward, or PyMC3. A task that was previously challenging due to technical issues handling the constraints. We also show some of the advantages of the new coordinates we introduce in our approach which includes better properties in formulation of priors, identifiability, and resulting posterioris that are less prone to multi-modality.

Preliminary work and draft manuscript.

1 Introduction

The Bayesian modeling paradigm involves setting up a probabilistic model describing how data was generated, assigning prior distributions over unknown model parameters, and then calculating a posterior distribution over these parameters [4; 17]. In practice, this posterior distribution is often intractable to compute exactly except for the simplest models. This often requires one to resort to approximate posterior inference algorithms based on Monte-Carlo sampling or Variational Inference (VI). Fortunately, a lot of progress has been made recently in this area including state-of-the-art algorithms such as Hamiltonian Monte Carlo sampling (HMC) [18], the No-U-Turn Sampler (NUTS) [9], Automatic Differentiation VI (ADVI) [11] and Black Box VI [20]. These algorithms are applicable to a wide class of models and are readily available in popular Probabilistic Programming languages such as Stan, Edward, and PyMC3 [3; 24; 21].

One class of models these algorithms do not generally apply to are models with parameters constrained to be unit-vectors or orthonormal matrices. This most notably precludes many models arising in several domains such as materials science [19], biology [7], and robotics [13], and also models arising in probabilistic dimensionality reduction [2; 10] such as PPCA, BXPCA [15], mixture of PPCA [5], CCA [17, Chapt. 12.5], and the examples we showcase in our empirical studies section. For simple constrained parameters, researchers and practitioners have typically bypassed this hurdle by transforming constrained model parameters to an unconstrained space and conducting inference in the resultant space. For example, if a model contains some parameter $\sigma > 0$ that is constrained to be positive, one can simply take the log of this parameter and conduct inference over $\tilde{\sigma} = \log \sigma$, which is unconstrained. This procedure is done routinely in Stan [3] and is the basis for ADVI [11].

Unfortunately, for more complex constraints, such transformations have not been mathematically derived or widely used. Many researchers have instead devised

various sampling algorithms for obtaining distributions over constrained parameters such as unit-vectors and orthonormal matrices [8; 1; 2; 10]. These algorithms use different update rules on constrained and unconstrained parameters often making them difficult to implement in standard software packages and precluding practical use on large complex models. In particular, no VI methods have been proposed for models with unit-vector and orthonormal matrix parameters making inference by practioners on large models with these parameters particularly problematic. Being able to use a transform would be ideal as it would more cleanly modularize models from inference algorithms.

To this end, we introduce an approach based on the Givens Transform. We provide a transform between the space of orthonormal matrices with constrained parameters to an unconstrained space. To the best of our knowledge, the use of Givens Transforms for this purpose has not been widely used before. Our approach greatly contributes toward the goal of allowing for the application of any general inference algorithm to models containing unit-vectors and orthonormal matrix parameters. Our approach is easy to implement and does not require any specialized inference algorithms or modifications to existing algorithms or software. This allows users to rapidly build and prototype complex probabilistic models with orthonormal matrix parameters in any common software framework such as Stan, Edward, or PyMC3 without having to worrying about messy implementation details. Users can then subsequently conduct fully Bayesian inference using any state-of-the-art inference algorithm available in these packages, including variational inference, which was previously challenging. We stress that allowing users to use models with orthonormal matrix parameters in common modeling packages opens up use of a wide class of new models, and frees them up to focus on modeling rather than implementation and debugging of custom modeling code and inference algorithms. Furthermore, by treating parameters agnostically as unconstrained, the Givens Transform allows inference algorithm designers to focus on more general algorithms rather than separate model specific ones.

In addition, our Givens Transform approach represents orthonormal matrices in terms of a sequence of fundamental rotations through given angles. This yields geometric insights into novel and useful ways to work with and interpret models with orthonormal matrix parameters. This helps in addressing a number of previously unresolved issues. Specifically, the elegant geometric representation lets us see how, by limiting the range of the parameters in the Givens Transform, we can naturally avoid issues of unidentifiability that arise when working with orthonormal matrices. The Givens Transform also enables new and creative ways to generate and use prior distributions on orthonormal matrices, and thus subspaces, a task that had previously been rather complicated due to the difficulty of evaluating densities of orthonormal matrix distributions for even small problem sizes [8]. As we shall discuss in more detail, our method allows for a natural way to specify prior distributions over orthonormal matrices comparable to the Matrix Langevin prior [16].

In Section 2 we discuss previous methods for conducting inference over unit-vector and orthonormal matrix parameters. We briefly explain how transformations are typically used in Bayesian inference in Section 3. In Section 4 we discuss the geometry of the Stiefel Manifold, the space of orthonormal matrices, setting the stage for our Givens Transform approach which we discuss in Section 5. In Section 6 we present several examples from our own applied work where we utilize the Givens Transform in Stan to implement several complex models containing unit-vector and orthonormal matrix parameters. We finish with a brief discussion in Section 7.

2 Related Work

A few sampling-based methods have been developed to obtain posteriors over orthonormal matrix parameters. Brubaker et al. [1] proposes the use of the SHAKE integrator [12] to simulate Hamiltonian dynamics and generate proposals. For constrained parameters, the integrator works by repeatedly taking a step forward that may be off the manifold using ordinary leap frog, then projecting back down to the nearest point on the manifold using Newton Iterations. Byrne and Girolami [2] as well as Holbrook et al. [10] exploit the fact that closed form solutions are known for the geodesic equations in the space of orthonormal matrices in the embedded coordinates, W. They utilize these equations to update constrained parameters in a different manner than for unconstrained parameters in their derived Embedded Manifold HMC (EMHMC) algorithm.

As these methods all use modified integrators for constrained parameters, they require additional bookkeeping of the support and the integrator of each model parameter, unlike the Givens Transform which treats these parameters as unconstrained parameters. This makes them incompatible with the current widely available Probabilistic Programming languages such as Stan and Edward, which typically do not expose the underlying inference algorithm to the user. Furthermore, these algorithms are unable to take advantage of improved samplers such as NUTS and optimization based approximate methods such as ADVI, limiting their scalability to large models.

3 Bayesian Inference of Constrained Parameters Using Transformations

We discuss in general how transformations can be used to work with constrained parameters corresponding to the Bayesian random variable Z with (possibly unnormalized) density $p_Z(z)$. One approach is to use a smooth locally invertible mapping $T : \text{support}(Z) \to \mathbb{R}^D$ to obtain a new density $p_U(u) = p_Z(T^{-1}(u))\,|J_{T^{-1}}(u)|$ in terms of an unconstrained random variable U. Here $J_{T^{-1}}(u)$ denotes the matrix Jacobian of T^{-1} of the coordinate transformation $u \to z$ and $|J_{T^{-1}}(u)|$ denotes its determinant. The determinant accounts for how a unit volume in u corresponds under the transformation with the volume in the original coordinates z to yield $\rho_Z(z)dz = \rho_U(u)du$ [4; 17; 11]. Under the transformation the density $p_U(u)$ can be used to obtain posterior samples $u_1, \cdots, u_N$ or a variational distribution $q_\gamma(u)$. An important property sought of the new coordinates is that the new parameters u are unconstrained but can be made to correspond to the original constrained parameters z of interest. Samples in $\mathbb{R}^D$ can then be freely mapped back to the original constrained space using the inverse transform T^{-1} to obtain posteriors in the original constrained space $z_1 = T^{-1}(u_1), \cdots, z_N = T^{-1}(u_N)$. While such a transformation may not always be possible in many cases of practical interest there exists such a map. We derive a transformation for orthonormal matrix parameters along these lines by appealing to the geometry of the space in which they reside.

4 Geometry of the Stiefel Manifold

We discuss some of the geometric properties of the space of orthonormal matrices. The most basic case consists of an orthonormal matrix with one column in $\mathbb{R}^n$ which corresponds to a unit vector. The collection of unit vectors form a sphere which is a sub-manifold of $\mathbb{R}^n$. Similarly, the collection of $n \times p$ orthonormal matrices correspond to a p-frame consisting of p orthonormal vectors that lie in an n-dimensional space. The collection of p-frames form a sub-manifold in the space of general $n \times p$ matrices. We refer to this space as the Stiefel Manifold and denote it by $V_{n,p}$ [16]. We can more formally define the Stiefel Manifold as

$$V_{n,p} := \{Y \in \mathbb{R}^{n\times p} : Y^T Y = I\}. \tag{1}$$

We remark that by convention all of the p-frames that comprise the Stiefel Manifold have the same orientation. Given a p-frame of $V_{n,p}$ with matrix representation Y, we can generate any other p-frame of $V_{n,p}$ by rigidly rotating the columns of Y around an appropriate combination of the axes.

Even though $n \times p$ orthonormal matrices are typically represented by np elements, the intrinsic dimension of the Stiefel Manifold, $V_{n,p}$, is actually $np-(p(p+1)/2)$. This arises from the constraints on the columns of the matrix that impose orthonormality. This dimensionality can be seen by observing that the first column of $Y \in V_{n,p}$ must have norm one and hence has one constraint placed on it. The second column must also have norm one and also must be orthogonal to the first column hence has two constraints placed on it. Continuing from the third column through the n^{th}, one arrives at the conclusion that each point of the Stiefel Manifold has number of degrees of freedom $np - (1 + 2 + \cdots + p) = np - (p(p+1)/2)$. The reduced dimensionality motivates using some type of rotation-based transforms which can be thought of as an $np-(p(p+1)/2)$-dimensional set of coordinates for elements of the Stiefel manifold.

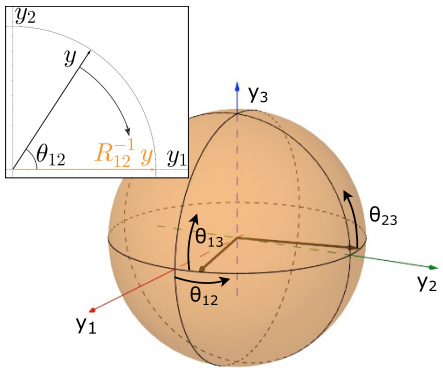

Figure 1: The Stiefel Manifold consists of all orthonormal p-frames. For $p = 1$ the Stiefel Manifold is equivalent to a sphere. We can represent the Stiefel Manifold for $p \geq 1$ as a successive composition of Givens rotations (inset) applied to a standard reference p-frame.

As a concrete example, the specific case where $n = 3$ and $p = 1$ corresponds to aforementioned unit vector in $\mathbb{R}^3$ whose position can be represented in terms of two angles of rotation: one representing rotation in the xy-plane, θ_{12} (latitude), and one representing rotation in the xz-plane θ_{13} (latitude). This is the standard spherical coordinates system with the radius free parameter set to exactly 1. Extending this to $n = 3$ and $p = 2$, we can imagine adding a second unit vector with position defined to be orthonormal to the first unit vector. However, now to define any other element of $V_{3,2}$ from any other, we must take care to keep the two orthonormal. This means we are constrained to rotate the second unit vector in reference to the first. Thus, this whole system is represented by three angles: two angles to represent the position of the first vector, and a third angle, θ_{23} that controls how much the sec-

ond basis vector is rotated about the first (Figure 1).

5 The Givens Transform

We represent $n \times p$ orthonormal matrices by a $np - (p(p+1)/2)$-dimensional vector of angles $\Theta := (\theta_{1,2} \cdots \theta_{1,n}) \cdots (\theta_{2,3} \cdots \theta_{2,n})(\theta_{p,p+1} \cdots \theta_{p,n})$. This corresponds to successively applying counter clockwise rotation matrices with these angles to the standard frame matrix $I_{n,p}$. We define $I_{n,p}$ to be the first p columns of the $n \times n$ identity matrix. More specifically, we represent an orthonormal matrix Y by

$$\begin{aligned} Y(\Theta) &= (R_{1,2}^{\theta_{1,2}} \cdots R_{1,n}^{\theta_{1,n}}) \cdots (R_{2,3}^{\theta_{2,3}} \cdots R_{2,n}^{\theta_{2,n}}) \\ &\quad (R_{p,p+1}^{\theta_{p,p+1}} \cdots R_{p,n}^{\theta_{p,n}}) I_{n,p}. \end{aligned} \tag{2}$$

We will choose by default that the elements of Θ be constrained to lie in either the interval $[-\pi, \pi)$ when tracking orientations or $[-\pi/2, \pi/2)$ when tracking only directors. We handle these constraints by applying a logistic transform element-wise similar to [4] to obtain a vector Φ of unconstrained values on $\mathbb{R}$. In particular, we use $\Theta = -L_1 + ((L_1 + L_2)/(1 + e^{-\Phi}))$ for $L_1 = L_2 = \pi$ or $L_1 = L_2 = \pi/2$ depending on the constraints on Θ. With these conventions we obtain using equation 2 our default represention Φ of an orthonormal matrix $Y = Y(\Theta(\Phi))$. We refer to this as approach as the Givens Transform and to simplify notation we denote it as $Y = Y(\Phi) = T^{-1}\Phi$ or $\Phi = \Phi(Y) = TY$ depending on direction considered. Our Givens Transform is invertible almost everywhere and provides a continuous map $\mathbb{R}^{np-(p(p+1)/2)} \longleftrightarrow V_{n,p}$ between the space of unconstrained parameters Φ and the space of orthonormal matrices Y of the Stiefel Manifold.

Our Givens Transform also can be viewed as a way to obtain and to compute coordinate charts for the Stiefel Manifold. Many other coordinate charts are possible and can be obtained by varying the angle conventions in the Given Rotations and ranges of the intervals. We emphasize that each of the angle-based coordinate charts have a set of singularities associated with them where the metric approaches zero. For example, in the case of $p = 1$ in $\mathbb{R}^3$ the Stiefel Manifold is a sphere and the angle coordinates correspond to the well-known Spherical Coordinates which have coordinate singularities at the poles. As with many differential geometry calculations on manifolds these singularities are not inherent to the Stiefel Manifold but rather an artifact of the coordinate description for the manifold. In practical calculations these singularities can be dealt with by changing coordinate charts as needed.

We also remark that one can further deal with singularities in practice by restricting the possible angles allowed to have non-zero probability density as a form of prior distribution over the data. This can be accomplished by using $L_1 = L_2 = \pi - \epsilon$ or $L_1 = L_2 = \pi/2 - \epsilon$ for a buffering value ϵ or by choosing a shift δ of the angles with $L_1 = \pi - \delta, L_2 = \pi + \delta$ or $L_1 = \pi/2 - \delta, L_2 = \pi/2 + \delta$. This could be useful not only for representing prior information but also for numerical purposes in calculations to avoid areas where the Givens Transform metric vanishes. By performing Bayesian inference with multiple shifts δ or restrictions ϵ one could still assess useful information about the relevance of these constraints and the full posteriori distribution.

Throughout the current paper we focus on the case when the posteriori distributions are sufficiently localized that regions near the singularities have negligable probability which allows us to perform calculations using only one coordinate chart. Since in Bayesian inference as more data is collected posteriori distributions tend to become more localized, our assumption amounts to dealing with inference scenarios with sufficient data to not need a global description of the Stiefel Manifold.

In the case of more broadly distributed posterioris requiring the use of multiple coordinate charts our methods still have utility. We can use the restriction approach above to still obtain useful information about the posteriori distribution. Also, for samplers the angle coordinates still offer advantages such as surpressing multi-modality or ridge concentration. For samplers one can obtain a common collection of samples from the samples $\{u_i^\alpha\}$ of the posteriori distribution considered in different charts α by using the pull-back to the embedding space of $n \times p$ matrices for the Stiefel Manifold as $\{z_i | z_i = T_\alpha^{-1}(u_i^\alpha), \alpha \in A\}$. Here, α denotes the index for the coordinate charts employed among some collection A. Analysis can then proceed as before with an appropriate local choice of the Givens Transform corresponding to a choice of local coordinate chart or by working directly in the embedding space.

While not pursued in the present work, similarly the Variational Inference (VI) methods could be treated reworking integration and computing likihoods using multiple Givens Transforms cooresponding to use of different local coordinate charts or using the transforms as a way to work in the embedding space. While we may purse these extensions in future work, we present here in this initial work the basic ideas for using Givens Transforms. We focus here primarily on the case when posteriori distributions are sufficiently localized to need description only by one coordinate chart for the Stiefel Manifold. This already allows for many models of interest to use orthonormal matrices in

their likelihood functions and to improve performance of algorithms for full Bayesian inference or Variational Inference as we present in the examples below.

We give a more formal specification of how our Givens Transforms are computed in Algorithm 1. We point out that our approach to obtaining the Givens Transform is closely related to the direct reduction algorithms from numerical analysis traditionally used for obtaining a QR factorization of an $n \times n$ matrix A [14].

Input: A set of $np - (p(p+1)/2)$ rotation angles θ.
Result: An orthonormal $n \times p$ matrix Y represented by the angles.
$Z = I_n$; idx $= 0$
for *i in 1:p* **do**
  for *j in i+1:n* **do**
    //create a counter clock-wise rotation matrix defined by theta
    $T = I_n$;
    $T[i,i] = \cos(\theta_{idx})$; $T[i,j] = \sin(\theta_{idx})$;
    $T[j,i] = -\sin(\theta_{idx})$; $T[j,j] = \cos(\theta_{idx})$;
    //apply the rotation matrix
    $Z = ZT$;
    $idx = idx + 1$;
  end
end
$Y = Z[:, :p]$;
Algorithm 1: Psuedo-code for obtaining the orthonormal matrix Y from the Givens Transform.

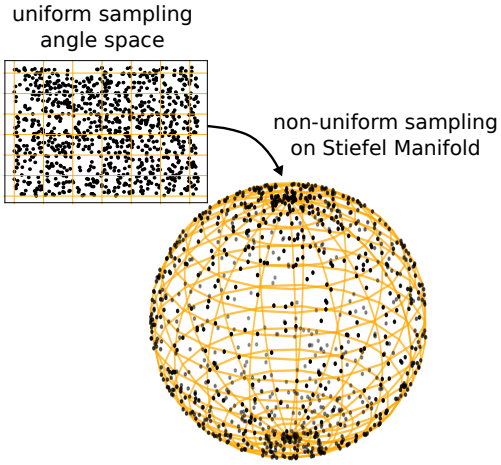

Figure 2: Importance of Jacobian Term. We show that if we sample uniformly in angle coordinates (inset) without the proper measure adjustment term how the distribution is non-uniform on the Stiefel Manifold when $p = 1$ (sphere). This illustrates the importance of the Jacobian term sometimes overlooked which accounts for how volumes are warped under the Givens Transform. In this case, uniform samples in angle space become sparse near the equator and congregate near the poles. Intuitively, this arises since under the mapping areas taken near the poles are shrunk far more than areas taken near to the equator. This results in points congregating closer to the poles of the sphere than the equator.

5.1 Transformation of Measure Under the Givens Transform

We discuss a few details important for Bayesian inference concerning calculation of the Jacobian adjustment factor for the measure accounting for the change in unit volume under the transformation as described in Section 3. The Jacobian adjustment term which is sometimes overlooked is important to ensure measures in the transformed coordinates correctly capture densities in the unconstrained parameter space (Figure 2).

For the Givens Transform computing directly the entries and determinant of the coordinate change matrix can be cumbersome. We present an alternative approach for this purpose. An $n \times p$ orthonormal matrix is np-dimensional and our Givens transform $\Phi(Y)$ maps this set to an $np-(p(p+1)/2)$-dimensional set of scalar parameters Φ that correspond to logistic transformed angles. To compute readily the Jacobian factor, we appeal to the exterior algebra associated with differential forms. The differential forms provide a convenient representation of the change in the infinitesimal volume form when switching from one space to the other. For accessibility, we also provide psuedocode for our mathematical expressions in the supplementary materials as well as an example Stan code.

For $n \times p$ orthonormal matrices, there are $np - (p(p+1)/2)$ free parameters and so the proper form to measure sets of orthonormal matrices is a $np-(p(p+1)/2)$-form. For an orthonormal, $n \times p$ matrix Y we can find an orthonormal $n \times n$ matrix G such that $G^T Y = I_{n,p}$. The matrix G in fact comes from the product of the appropriate rotation matrices that arises in the Givens Reduction Q in QR factorization. One can show as in Muirhead [16] that the adjustment term for measuring volumes on the Stiefel Manifold can be expressed in terms of the wedge product of differential forms. This corresponds to elements of the $n \times p$ matrix $G^T dY$ were we use terms that lie below the diagonal. This can be expressed in terms of the wedge product on differential forms as

$$S = \bigwedge_{i=1}^{p} \bigwedge_{j=i+1}^{n} G_j^T \, dY_i \tag{3}$$

where S gives the volume form expressed for entries of the orthonormal matrix Y. The $\wedge$ notation de-

notes for two differential forms α and β the differential form $\gamma = \alpha \wedge \beta$. The G_j is the j^{th} column of G and Y_i is the i^{th} column of Y. To obtain the form in angle coordinates, we use dY_i in terms of the angle coordinates by the following relationship $dY_i = J_{Y_i}(\Theta)\, d\Theta$, where J_{Y_i} is the Jacobian of Y_i with respect to the angle coordinates. Once we obtain the form (3) in terms of the angle coordinates the result is a wedge product of $np - (p(p+1)/2)$ co-vectors that span a $np - (p(p+1)/2)$ dimensional space. This can be reduced to the determinant by lowering indicices to obtain vectors which are aligned side-by-side as a $np - (p(p+1)/2) \times np - (p(p+1)/2)$ matrix. This determinant gives the Jacobian adjustment for the probability densities under the Givens Transformation.

We further remark that we can insert this into the log-probability of a model to avoid the sort of unintended sampling behavior depicted in Figure 2. We use the Jacobian adjustment term of equation 3 when computing the log-probability in all of our reported examples and results.

6 Results and Examples

To demonstrate the use of our Givens Transform approach, we construct several models with orthonormal matrices or unit vectors as parameters and perform fully Bayesian inference.

6.1 Avoiding Unidentifiability in Neural Network Models Using Unit-Vectors

A single layer neural network with Rectified Linear Unit (ReLU) non-linearities and H hidden nodes maps inputs $x \in \mathbb{R}^D$ to outputs $y \in \mathbb{R}$ via the relationship $y = hW_2 = \max\{0, W_1 x + b_1\}W_2$ where $h = \max\{0, W_1 x + b_1\} \in \mathbb{R}^H$, $W_1 \in \mathbb{R}^{H \times D}$ and $b_1, W_2 \in \mathbb{R}^H$. It can readily be seen that several equivalent values of W_1, b_1, and W_2 result in the same exact function. For example, the ith row of W_1 and the ith entry of b_1 can be scaled by α as long as the ith entry of W_2 is scaled by $1/\alpha$ [6, p.277].

In Bayesian analysis, this leads to thin, ridge-like posterior distributions with high curvatures that are difficult to sample and approximate using Variational Inference (VI). As a consequence this often times leads to variational posteriors that underestimate the model uncertainty (Figure 3). We can obtain much more well-behaved posteriors by replacing the columns of W_1 with unit-vectors instead. We use our Givens Transform approach to sample over the space of unconstrained angles, see Figure 3. We demonstrate also this idea in a few other related settings.

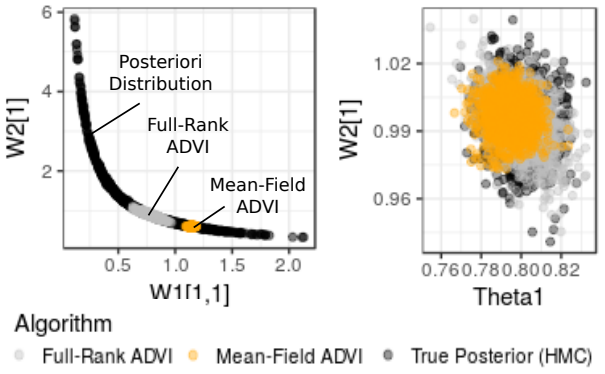

Figure 3: On the left we show the Posterior HMC samples of $W_1[1,1]$ and $W_2[1]$ to reveal how the scaling unidentifiability of ReLU networks manifests as ridge-like posteriors that are difficult to sample and approximate using Variational Inference (VI). On the right we show the posteriors using the coordinates associated with our Givens Transform representation which are much more well-behaved showing good approximations using Variational Inference (VI).

6.2 Probabilistic PCA (PPCA)

Factor Analysis (FA) and Probabilistic PCA (PPCA) [23] posit a probabilistic generative model where high-dimensional data is determined by a linear function of some low-dimensional latent state [17, Chapt. 12]. Geometrically, for a three-dimensional set of points forming a flat pancake-like cloud, PCA can be thought of as finding the best 2-frame that aligns with this cloud (Figure 4). Formally, PPCA posits the following generative process for how a sequence of high-dimensional data vectors $\mathbf{x}_i \in \mathbb{R}^n$, $i = 1, \cdots, N$ arise from some low dimensional latent representations $\mathbf{z}_i \in \mathbb{R}^p$ $(p < n)$. PPCA posits the data can be represented by a linear transformation represented by matrix $W \in \mathbb{R}^{n \times p}$ with

$$\begin{aligned} p(\mathbf{z}_i) &\sim \mathcal{N}_p(0, I) \\ p(\mathbf{x}_i | \mathbf{z}_i, W, \sigma^2) &\sim \mathcal{N}_n(W\mathbf{z}_i, \sigma^2 I). \end{aligned} \tag{4}$$

A closed-form maximum likelihood estimator for W is known for this model in the limit as $\sigma^2 \to 0$, but as we shall see, for more complicated models/likelihoods, closed-form maximum-likelihood estimators are almost never known. This has often been dealt with by using Expectation Maximization (EM) in these models to obtain a point estimate [17, Chapt. 12.2.5]. In Bayesian inference we are typically interested in the entire distribution over possible solutions, i.e. a posterior distribution over unknown parameters to quantify uncertainty.

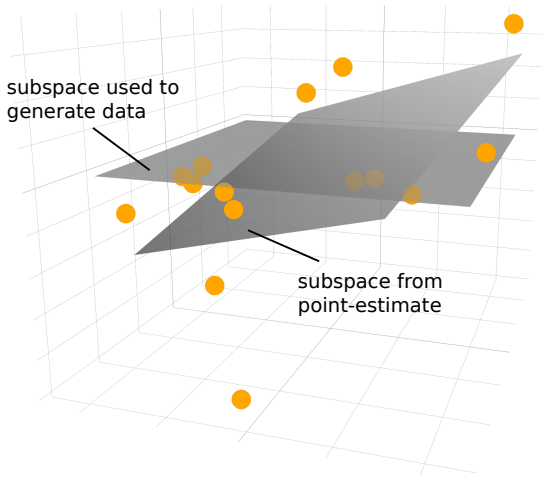

Figure 4: PCA finds a single orthonormal matrix in the Stiefel Manifold that best describes the subspace that the data lie in. This point estimate can often mislead us from the true subspace which in this case is the horizontal xy-plane which was used to generate the noisey data. Alternatively, in Probabilistic PCA (PPCA) a posterior distribution is used to estimate the approximating subspace and also to quantify uncertainty of the result. To obtain tractable methods requires a good representation for the latent variables in PPCA.

One can show that the W parameter in PPCA is unidentifiable [17, chapt. 12.1.3], as it can be rotated to achieve an identical likelihood; thus the model must be changed to make the posterior distribution interpretable. Furthermore, this rotational unidentifiability manifests in the log-likelihood function as regions in parameter space where large curvature arise, causing numerical problems in HMC, as pointed out by Holbrook et al. [10]. To this end, those in the Bayesian dimensionality reduction community have used a modified form of the model (4), whereby the matrix W is replaced by a new term $W\Lambda$ where W is an $n \times p$ orthonormal matrix and Λ is a $p \times p$ diagonal matrix with positive elements [2; 10].

6.2.1 Test on Synthetic Data

We used the Givens Transform to fit this modified PPCA using Stan's NUTS inference algorithm, which otherwise would be unusable on this model with an orthonormal matrix parameter. We generated a synthetic, three-dimensional dataset that lies on a two-dimensional plane with $N = 15$ observations according to the modified version of (4) (data shown in Figure 4).

We choose $\text{diag}(\Lambda) = \text{diag}(1, 1)$, $\sigma^2 = 1$, and W to be $I_{3,2}$, which in the Givens representation corresponds to $\theta_{12} = \theta_{13} = \theta_{23} = 0$ i.e. the horizontal plane, which contrasts with the slanted plane that we obtain from a classical PCA maximum likelihood estimate (Figure 4). In this case the advantage of the full posterior estimate the Bayesian framework affords is clear. Posterior samples of θ_{13}, which if we recall from Figure 1 is the Givens Transform angle that controls the upwards tilt of the plane, reveal a wide posterior which cautions us against the spurious maximum likelihood estimate of $\hat{\theta}_{13} = -0.15$ (Figure 5).

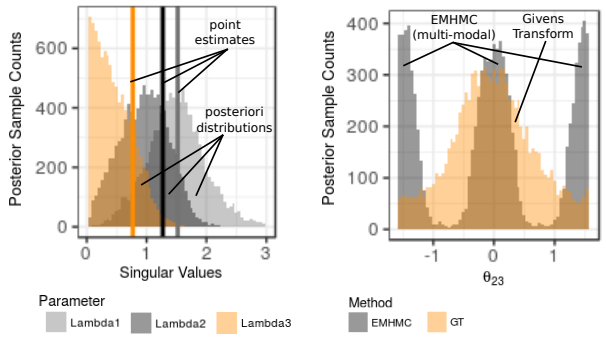

Figure 5: PPCA inference for three-dimensional synthetic data. (Left) Posterior draws of the Λ parameter are more informative in dimensionality selection than point-estimates. (Right) By limiting the angles of rotation in the Givens Transform we can further avoid unidentifiability in our problem and eliminate multimodal posteriors that show up in other methods such as EMHMC.

We also note that the representation of the Givens Transform can in certain models allow us to avoid issues of identifiability that are present in the sampling algorithms of [1; 2]. While most identifiability issues are alleviated by using the modified PPCA with an orthonormal matrix, the PPCA likelihood is still equivalent for an orthonormal matrix W and any permutation of the columns of W which reverse orientations (negative permutations) [17; 10, Chapt. 12.1.3]. Taking a geometric view of the Stiefel Manifold (Figure 1), this means that a mirroring of the p-frame would yield an identical value in the likelihood of even the modified PPCA. As such, even the methods of Brubaker et al. [1] and Byrne and Girolami [2] will lead to multimodal posteriors that can be avoided in a straightforward manner by simply limiting the angles in the Givens Transform from a range of $(-\pi, \pi)$ to a range of $(-\pi/2, \pi/2)$, a change that is much more evidently afforded when working in the angle coordinates (Figure

5 [right]).

We remark about the role of coordinate charts and degeneracy of the associated metric when the true subspace basis lies in the Stiefel Manifold near the analogue for a sphere of the "pole," i.e. when θ_{ij} is close to $-\pi/2$ or $\pi/2$. In this case the posteriors might still tend to be multi-modal as the region in parameter space close to these boundaries may be nearly equal. Ideally, in calculations these regions should contain little probability mass. In these cases, when the posteriors are sufficiently localized one can simply change the coordinate bounds (chart) to deal with this issue so that $\theta_{ij} \in (0, \pi)$ will have a unimodal posterior in the new coordinate system. This also can be used to alleviate possible exploration issues in HMC where the posterior may exhibit artifical small mass density or other inaccuracies in such a region. In Stan this is straightforward, as one simply has to change the lower and upper bound of the angle parameter (which through the logistic transform defines our unconstrained parameters and hence chart).

6.3 Hierarchical subspace models for grouped multi-view medical data

We modeled grouped multi-view hospital data for injured patients using a hierarchical CCA model [17, Chapt. 15.2]. CCA can model two types (or views) of data as being a function of two respective latent low dimensional states, but also a common latent state that captures the common information contained in both views (Figure 6 [left]). In our case we compared blood protein measurements and clot strength measurements for injured patients belonging to one of four groups, depending on the type of injury. While the four types of injuries were different enough so that we could not use a single CCA model to capture the characteristics of all models at once, the four groups were not so different as to warrant separate CCA models for each. To share information between the CCA models, we placed a hierarchical prior over the angles of the Givens Transform representing the distinct orthonormal matrix parameter for each group.

While distributions on the Stiefel Manifold such as the Matrix Langevin distribution [16] exist, these distributions are difficult to use in practice, as computing their density requires evaluating an expensive matrix sum [8]. By appealing to the Givens Transform and placing a hierarchical prior over the angles of the different orthonormal matrices, we were able to build a hierarchical model over subspaces, a previously intractable task. The hierachical prior "shrinks" the posterior median of the orthonormal matrices towards a common mean in addition to reducing the variance of these estimates (Figure 7). This is particularly helpful for groups with only a smaller number of observations such as the SW group, which contains only 16 patients, in comparison with the GSW group of 86 patients. Comparing the angle between the first principal components for the SW and GSW groups illustrates how using a hierarchical prior shrinks estimates of subspaces together towards a common hierarchical subspace (Figure 8).

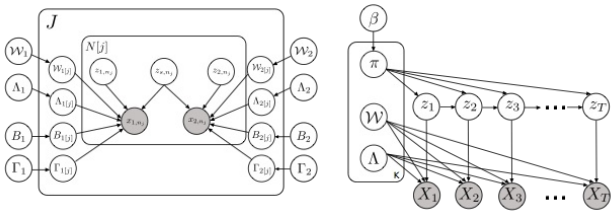

Figure 6: Probabilistic graphical models for Hierarchical CCA Model (left) and Network HMM (right).

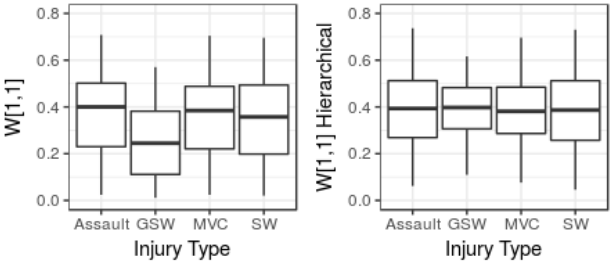

Figure 7: (Left) When estimated separately, estimates of the matrix parameter W have high uncertainty. (Right) Placing a hierarchical prior over these matrices with GT-PPCA shrinks these parameters to a common hierarchical mean and results in smaller posterior intervals.

6.4 Social Networks

We built an HMM subspace model for count data to model the hidden time-dependent structure of a social network of school children. RF sensors were used to track the interactions between school children in 12 different classes (two classes for grades 1-6) for an entire school day so as to better understand how disease spreads throughout a network [22]. We collated the number of interactions between each pair of classes into 11-minute contiguous time windows, giving us 177 symmetric matrices of counts representing the network interactions between different classrooms throughout the day (Figure 9 [lower row]). We modeled the ele-

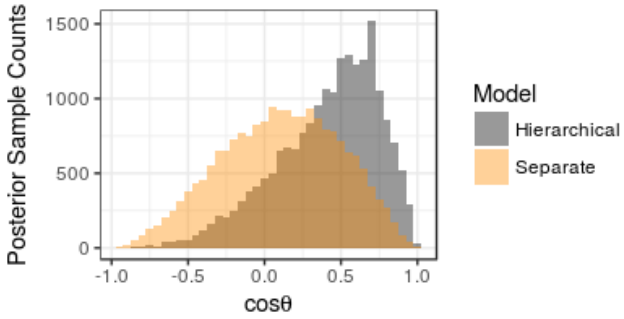

Figure 8: Geometrically the respective first principal components of two different groups are shrunk closer together in a hierarchical model.

ments of these count matrices as each coming from a Poisson distribution with rate defined by a symmetric matrix $R = \exp(W\Lambda W^T)$, where the orthonormal matrix W captures the low-dimensional structure of the network. To model the time varying structure of the network, we posited that the network was always in one of three latent states, that evolve according to a Markov Chain (Figure 6 [right]). The three states each have their own associated orthonormal matrix W_i that captures the low-dimensional latent network structure for that state.

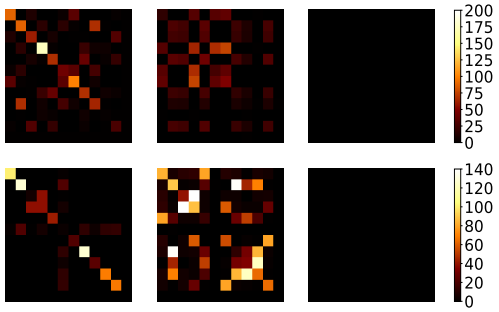

Figure 9: Posterior modes of rate matrices for the three states (top) capture the pattern found in example count matrices belonging to each of these three states (bottom).

The posterior modes capture the latent structure of the rate matrices of the three hidden states (Figure 9 top row). Interaction rates are visibly low when students are in class, high during lunch, and nonexistent when out of school out of class. Posteriors of the orthonormal components of the rate matrices are shown in (Figure 10 [left]). We also generated samples from the posterior distribution over states from posterior samples of the Markov Chain, enabling us

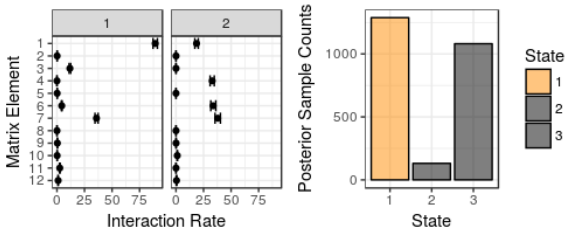

Figure 10: (Left) Posterior intervals from GT-PPCA with NUTS capture uncertainty in the orthonormal matrix estimates for the first two columns of the rate matrix for the first hidden state. (Right) Posterior draws can tell us the posterior probability that the network was in a certain state given the data.

to provide a posterior over which of the hidden states the network is in at a given time (Figure 10 [right]), a common inference task in disease networks as well as fMRI networks.

7 Discussion

We introduced an approach for describing orthonormal matrices using unconstrained parameters based on the Givens Transform. Our approach facilitates the construction and inference of complex probabilistic models with unit-vector and orthonormal matrix parameters. We show using real-world examples how one can use Stan and our Givens Transform approach to obtain uncertainties over such parameters. Our approach also is helpful in reducing multi-modal posteriors allowing for analyzing parameters in terms of an often easier to understand angle representation. We expect our approach can be used quite widely for the study of probablistic models benefitting from orthonormal matrix representations.

Acknowledgements

The author P.J.A acknowledges support from research grant DOE ASCR CM4 DE-SC0009254 and NSF DMS - 1616353.

References

[1] M. Brubaker, M. Salzmann, and R. Urtasun. A family of mcmc methods on implicitly defined manifolds. In *Artificial Intelligence and Statistics*, pages 161–172, 2012.

[2] S. Byrne and M. Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

[3] B. Carpenter, A. Gelman, M. Hoffman, D. Lee,

B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 20, 2016.

[4] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[5] Z. Ghahramani, G. E. Hinton, et al. The em algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto, 1996.

[6] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[7] T. Hamelryck, J. T. Kent, and A. Krogh. Sampling realistic protein conformations using local structural bias. *PLoS Computational Biology*, 2 (9):e131, 2006.

[8] P. D. Hoff. Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2): 438–456, 2009.

[9] M. D. Hoffman and A. Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[10] A. Holbrook, A. Vandenberg-Rodes, and B. Shahbaba. Bayesian inference on matrix manifolds for linear dimensionality reduction. *arXiv preprint arXiv:1606.04478*, 2016.

[11] A. Kucukelbir, R. Ranganath, A. Gelman, and D. Blei. Fully automatic variational inference of differentiable probability models. In *NIPS Workshop on Probabilistic Programming*, 2014.

[12] B. Leimkuhler and S. Reich. *Simulating hamiltonian dynamics*, volume 14. Cambridge University Press, 2004.

[13] F. Lu and E. Milios. Robot pose estimation in unknown environments by matching 2d range scans. *Journal of Intelligent and Robotic systems*, 18(3): 249–275, 1997.

[14] C. D. Meyer. *Matrix analysis and applied linear algebra*, volume 2. Siam, 2000.

[15] S. Mohamed, Z. Ghahramani, and K. A. Heller. Bayesian exponential family pca. In *Advances in neural information processing systems*, pages 1089–1096, 2009.

[16] R. J. Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009.

[17] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[18] R. M. Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2 (11), 2011.

[19] S.-H. Oh, L. Staveley-Smith, K. Spekkens, P. Kamphuis, and B. S. Koribalski. 2d bayesian automated tilted-ring fitting of disk galaxies in large hi galaxy surveys: 2dbat. *Monthly Notices of the Royal Astronomical Society*, 2017.

[20] R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

[21] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016.

[22] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one*, 6(8): e23176, 2011.

[23] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[24] D. Tran, A. Kucukelbir, A. B. Dieng, M. Rudolph, D. Liang, and D. M. Blei. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*, 2016.